DOCUMENT RESUME

ED 167 590 TM 008 347

AUTHOR Donlon, Thomas F.; Fitzpatrick, Anne R.
TITLE The Statistical Structure of Multiple-Choice Items.
PUB DATE [Oct 78]
NOTE 33p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, New York, October, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Complexity Level; *Confidence Testing; Guessing (Tests); *Item Analysis; Multiple Choice Tests; *Scores; *Scoring Formulas; Senior High Schools; Test Construction; Test Interpretation; *Weighted Scores

ABSTRACT

On the basis of past research efforts to improve multiple-choice test information through differential weighting of responses to wrong answers (distractors), two statistical indices are developed. Each describes the properties of response distributions across the options of an item. Jaspen's polyserial generalization of the biserial correlation coefficient is suggested as an index of relationship between item distractors and a criterion, and a modified form of Shannon's uncertainty measure is proposed to assess the information potential of the frequency distribution across the incorrect item options. It is suggested that these indices be used to determine whether informational gains are likely to be achieved through differential weighting of item distractors. It is also suggested that they be applied to the distributions of non-right responses only, to remove the effect of the basic measurement potential of the items when dichotomously scored. Empirical results are presented which compare these indices to the difficulty and discrimination parameters of traditional item analysis, using response data from verbal and mathematical subtests of the Scholastic Aptitude Test. These item-type comparisons are used to evaluate the analytic effectiveness of the proposed statistics. Applications to further work in differential response weighting are discussed. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED167590

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Statistical Structure of Multiple-Choice Items

Thomas F. Donlon

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Thomas F. Donlon

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

and

Anne R. Fitzpatrick

University of Massachusetts - Amherst

NERA

Ellenville, N.Y.

1978

IM008 347

ABSTRACT

A commonly held premise is that the incorrect responses to multiple-choice items may be sufficiently informed by partial knowledge to comprise a valuable source of test information and should be accessed by differentially weighting and scoring these wrong response options. This premise, however, is seldom tested systematically at the item level. The specific distributional properties of responses to wrong answers, or distractors, in multiple-choice items subjected to weighting are seldom reported.

This paper reviews past research efforts to improve test information through differential weighting and identifies, on the basis of this research and from theoretical considerations, two statistical indices which can be used to describe the properties of response distributions across the options of an item. Jaspen's (1942) polyserial generalization of the biserial correlation coefficient is suggested as an index of relationship between item distractors and a criterion, and a modified form of Shannon's (1948) uncertainty measure is proposed to assess the information potential of the frequency distribution across the incorrect item options. It is proposed that these indices be used to determine whether informational gains are likely to be achieved through differential weighting of item distractors. It is suggested that they be applied to the distributions of non-right responses only, to remove the effect of the basic measurement potential of the items when dichotomously scored.

Empirical results are presented which compare these indices to the difficulty and discrimination parameters of traditional item analysis, using response data from verbal and mathematical subtests of the College Board SAT. These item-type comparisons are used to evaluate the analytic effectiveness of the proposed statistics. Finally, the potential application of these indices to further work in differential response weighting and to research in other areas is explored.

The Statistical Structure of Multiple-Choice Items

Thomas F. Donlon
Educational Testing Service

and

Anne R. Fitzpatrick
University of Massachusetts - Amherst

There have been numerous attempts to draw more information from multiple-choice measures than is to be obtained through simple dichotomous scoring. An extensive review of these efforts is provided by Stanley and Wang (1970), under the general label "differential weighting." Early work involved weighting the correct responses to items. Underlying the work in this area by researchers such as West (1924), Odell (1931) and many others was the reasonable supposition that items comprising a test might contribute differentially to the overall reliability or validity of that test. If this were the case, the conclusion was warranted that the equal item weights implied by the traditional method of summing or averaging individuals' correct responses should be abandoned in favor of item weights which varied over items and reflected the informational value of each item.

Although this reasoning seemed sound, Wang and Stanley (1970) were led to observe "numerous studies have been performed demonstrating that item weighting is futile." Some reasons for this were developed in formulations by Wilks (1938), and Gulliksen (1950), which clearly indicated that a correlation between weighted and unweighted measures comprised of dichotomously scored items rapidly approaches unity if the component items of these measures

are numerous and positively inter-correlated. Unless a test was composed of only a few independent items, no set of item weights would confer more informational advantage than that obtained by simply assigning each item a unit weight. This highly correlated relationship between weighted and unweighted sets of items was demonstrated empirically by Stalnaker (1938), among others, and explains the frequent finding that unit weights contribute as effectively as variable item weights to test reliability and validity (Guilford, et al, 1942; Fruchter, 1962). As suggested by Wang and Stanley (1970), it appears that only more complex approaches to item weighting, such as Birnbaum's (1968) three parameter latent trait model, which specifies item weights that vary over levels of ability, Samejima's (1969) grades response model, and Cleary's (1960) weighting of test items over individuals, have promising if possibly restricted usefulness.

Wang and Stanley saw more promise, however, for what they called "variable response weighting methods," in which weights are assigned to each option in a multiple-choice test.

The scoring of incorrect item options has been explored empirically. Weights for wrong response options are derived by a priori or empirical means, while unit weights for correct response options are, in some instances, retained. Formula scoring techniques, for example, often used to adjust total scores for chance success, establish on an a priori basis different weights for wrong and/or omitted responses. Empirical methods, on the other hand, often apply multiple regression techniques to identify optimal weights which will maximize the correlation between the combination of total wrong and total right scores and a criterion measure (Kelley, 1923; Thurstone, 1919).

Guilford (1950) has reported that gains in validity from such an optimal weighting procedure usually range from .03 to .06 and exceed those resulting from the simple formula scoring method. While these gains are slight, they involve the use of a constant weight for the wrong responses and can not be considered a fully differential weighting scheme, as various distractors are not assigned different weights.

Differential option weighting using a priori methods has been initiated by those who make what Lord and Novick (1968) call a "realistic" assumption that partial knowledge may be represented in examinees choices of distractors. In this circumstance, scoring distractors, in terms of their degrees of correctness, could be expected to produce more precise estimates of the measured abilities or better correlations with a criterion measure. Aiken (1967) provided a theoretical demonstration of the effective increase in total test score variance to be had by assigning distractors different weights and thereby provided some theoretical basis for the procedure of option weighting. Empirical studies of differential option weighting suggest that a portion of this increase may possibly be true score variance, although more study of total test score variance would be required to confirm that possibility (Wang & Stanley, 1970).

To establish a priori weights for distractors, the two main approaches have used expert judges and empirical data separately or in combination. Expert ratings are most frequently used to establish an order for existing response alternatives, a step which is post facto to test development, as distractors are rarely designed to assume a hierarchy of correctness. Nedelsky (1954) showed that the reliability of a physical science achievement

test could be somewhat improved throughout the ability range if total scores on the test were corrected for respondents' choices of those options judged by experts to be grossly incorrect. Nedelsky's results also showed that the achievement of low ability students could be estimated with significantly more accuracy if the total scores of these students were derived from a tally of their choices of the very incorrect distractors alone. Jacobs and Vandeventer (1970), working with the Raven's Progressive Matrices, used facet analysis as a basis for judged similarity of wrong answers to correct answers, identifying superior or partially correct distractors in 18 items and half of the selected distractors correlated positively with the total scores on that test.

The assignment of option weights by empirical methods have similarly produced some favorable findings. To compare the usefulness of formula scoring and weighted option scoring, Hendrickson (1970) derived response weights for the items of two verbal and two mathematical subtests of the SAT using a technique described by Guttman (1941) for maximizing internal consistency reliability. In accord with Guttman's technique, she weighted each response option of an item by values linearly related to the total scores on the remaining items of respondants who chose that option. Iterative procedures and subsequent cross-validation of the weights produced option weights which contributed to an average increment in the reliabilities of the four subtests equivalent to that which would result if these tests were lengthened by 49%. Using a principal components variant of the Guttman approach, Serlin & Kaiser (1978) found a maximum alpha value of .7736 in scoring a very short 5-item test, considerably above the coefficient value

of .4367 attained by scoring the items dichotomously. Davis's (1959) demonstration that coefficients of correlation between option and total scores provided acceptable estimates of Guttman weights stimulated Davis & Fifer's (1959) use of correlation coefficients modified by expert judgments as option weights to increment the parallel forms reliability of a mathematical reasoning test. These authors found reliability improved, by this technique, to a level otherwise obtained only if the test were lengthened one and a half times. Furthermore, no lowering of concurrent validity was reported, even though distractor weights employed in their validity study were selected for reliability-maximizing purposes. Sabers & White (1960) also used the correlational approach to distractor weighting in the interest of improving test validity, but found both validity and reliability unresponsive to the weighting efforts.

It is important to note, in evaluating such research, the likely and problematic incompatibility between weights designed to maximize reliability and those derived to maximize validity. To maximize reliability, the most heavily weighted options are likely to be those which increment internal consistency; homogeneous items will, therefore, be preferentially weighted. In contrast, a set of validity-maximizing weights will favor heterogeneous items. These two sets of weights will often not be the same. Although Davis and Fifer did not forfeit validity when their reliability-maximizing weights were retained in their study of validity, Hendrickson (1970) presents some findings which suggest that Guttman's procedures for enhancing internal consistency will be costly to validity. Lord & Novick (1968) demonstrates the theoretical basis for this occurrence.

A study reported by Thissen (1976) proves to be perhaps the most encouraging to the enterprise of differential option weighting. His application of a multiple category latent trait model produced evidence which provides strong support for the assumption that wrong response options can operate to distinguish degrees of partial knowledge when their underlying structure permits, and may provide access to more information than is available from conventional scoring procedures. Noting Jacob & Vandeventer's findings of frequent relationships between total scores on the Raven Progressive Matrices measure and distractors of that test, Thissen classified all response alternatives from parts of this test in terms of their degree of correctness. Thissen found that, for the lower half of the ability scale, the information yield of multiple response scoring exceeded that of dichotomous scoring by factors of one-third to nearly two. In addition, Thissen reported that the characteristic curves for the most to least popular response options of each item indicated that these options discriminated differentially, with distractors contributing more information than the right-answer option at the lower end of the ability scale. Greater discrimination, and therefore more information about abilities was indicated, in fact, for items varying substantially in their power to discriminate when dichotomously scored. Total test information functions, which trace changes in the reciprocal of measurement error over levels of ability, reiterated the considerable gains in information yielded in the lower half of the ability spectrum through application of multiple category rather than binary scoring procedures. Thissen suggested that such gains would not be reaped were the distractors not constructed with differing,yet known relationships to ability. As suggested by Hambleton et al (1972), it is the

respondents of middle and lower abilities who will make choices among the incorrect response options, so that greater precision is likely to accrue in estimates of average and lower abilities.

Although results favorable to the weighting of incorrect response options have emerged and can not be ascribed to chance, only the measures used by Thissen were apparently designed to accommodate differentially weighted wrong response options. The distractors of the achievement items used by Nedelsky, those employed by Davis and Fifer, and the SAT items utilized by Hendrickson were chosen to satisfy only the traditional criteria of item quality, criteria established primarily in connection with dichotomously scored items. In conventional analysis, the merits of distractors are not judged by a criterion of differential attractiveness. As noted by Henrysson (1971) a distractor is usually acceptable simply if it appeals to substantial numbers of examinees who represent the low and middle rather than the upper ranges of the ability scale. Occasionally it is recommended that the incorrect alternatives be chosen by equal percentages of those who don't know the right answers and some experts suggest that good distractors are those which are equally attractive (Guilford, 1942; Gronlund, 1976), so that choices among them will be essentially random and simple formula score weights, therefore, appropriate to apply.

Yet as Davis suggested in 1959, only if the various distractors of a set of items are structured to attract respondents of differing abilities will differential scoring of the options add more information than is accessible through binary scoring. The results of the research in differential option weighting imply that distractors constructed even by traditional methods do evince a definite, if unintended, capacity for discriminating

between levels of abilities. Were the wrong response options of test items explicitly selected for their capacity to exhibit a structure of varying correctness, better results could be anticipated. As Davis concisely states:

> "The measurement of variance attributable to selection among incorrect choices on the part of examinees who cannot identify the correct choices to the items requires, first, that the incorrect choices in each item actually represent, with respect to the variable measured, different levels of partial information of skill and of partial misinformation or ineptitude."

We would propose, with Davis, that the information yield effected by differentially weighting incorrect options is to be maximized when these options represent varying degrees of correctness and can be ordered to reflect different positions along an ability dimension. In fact, we believe that the weighting of tests should be attempted only when the items are believed to have useful levels of distinctiveness and orderliness. To evaluate these characteristics we propose the application of two indices: 1) the multiserial correlation suggested by Jaspen (1946) and the Relative Uncertainty measure suggested by Pike and Flaugher (1970). There are, of course, other and related indices which could serve these functions. No systematic evaluation of these other techniques has been attempted. The general proposition is simply that the statistical characteristics of the items to be used in studies of differential weighting are too often neither reported nor computed.

Description of the indices

Studies of the effects of differential weighting are typically attempted with tests which were, in fact, constructed with more conventional scoring methods in mind. The typical inquiry attempts to determine whether the more complex differential scoring technique will be more powerful in its extraction of information latent in the patterns of wrong answers, omissions, or failures to complete. To the knowledge of the authors, no test has been specifically constructed with predetermined optimal item properties for weighting. The repeated demonstrations of failures, then, with weighting schemes on ad hoc material, fail to advance our understanding of the possible improvements to be attained through differential weighting if the average potential of the individual test items were greater. All too often the justification for an attempted weighting is the exhibition of a single test item for which there is a strong distracter sufficiently competitive with the correct response to define a "second-best" group of respondents. The promise of weighting lies in the prospect of capturing such situations, but the extent to which they exist in sufficient number to be useful is in effect left to be implicitly tested in the assessment of the score properties. The indices proposed in this paper permit an explicit testing of the item properties.

A number of indices could be proposed. The weighting enterprise is sufficiently sensitive to a diverse set of factors that quite powerful statistical methodologies could be advanced. The polyserial correlation and the relative uncertainty measure, however, are essentially the analogs, respectively, of the item discrimination index and the item difficulty index of traditional test development. They are proposed as useful, practical

approaches to the assessment of item properties. The merits of optional alternate approaches such as point-multiserial (Das Gupta, 1960), multiserial eta (Wherry and Taylor, 1946), and of Chi-square tests of distributional properties has not been assessed here.

The function of such indices in test construction for weighting is to insure that the item distracters provide information which is in agreement with the criterion and that there is sufficient information to make the extra scoring effort worthwhile. Traditional test construction tends to rely on an index of item-criterion relationship such as the biserial or point-biserial correlation coefficient to insure that the item,when scored,will agreed with the criterion and with other items. Traditional test construction tends to rely on an index of percent passing to gauge the overall information utility of the item. Basically, this index is interpretable in terms of where on the ability continuum the item will offer the most effective discrimination. But it is well known that, in terms of information yielding items of middle difficulty, with a proportion passing of .50, offer the most information, at least in the sense of the greatest number of discriminations among individuals. Extremely difficult or extremely easy items are rejected by test developers because they do not offer enough useful information, no matter how well correlated they are with the criterion, in most applications. Most traditional tests are evaluated through their potential for a broad-gauge assessment across a fairly wide span of abilities. Most studies of differential weighting appear to share this focus. To be useful for weighting, then, the response

categories to items must yield sufficient information, in the intuitive sense of discriminations among individuals, to make a detectable contribution.

The formulas for calculating Jaspen's polyserial correlation coefficient are provided in Appendix A. Initially it was hoped that the polyserial correlation coefficients could be calculated for what were called "whole item" distributions: distributions which included the correct answer category. Using such "whole item" data, the test of potential usefulness for weighting would lie in the extent to which the "whole item" polyserial was greater than the biserial coefficient computed for the same item.

In practice, with the actual items studied, this proved ineffective because the polyserials were too often constrained to similarity with the biserials through the extensive redundancy, and overlap in the information they presented and because a high polyserial could ambiguously reflect either a useful set of distracters or a basically effective right answer. Accordingly, the index applied was r_{pw}, the polyserial correlation of the wrong answers. This computation demands a number of Procrustean assumptions in order to apply Jaspen's algorithm, of which the strongest are clearly that the criterion measure remains approximately normal for the distribution of non-right respondents, and that the variance of these criterion measures is essentially that of the total group. The effect of those assumptions, and the usefulness of alternate statistical strategies remains to be tested. Since the primary interest in the study lies in relative comparisons of items, and since the effects of the assumptions will not much alter these relative standings, the general logic of the inquiry is not affected. The two

assumptions are routinely made in item analysis procedure at ETS when confronting the problem of computing item analysis data for groups affected by dropout.

The Relative Uncertainty measure was proposed by Pike and Flaugher and applied by them to tests of randomness in responses of minority children to test items. In applying Shannon's (1948) measure of average uncertainty, U, they noted that it has a different theoretical maximum for different numbers of response categories. Wishing to compare test items with different numbers of responses, Pike and Flaugher reported Relative Uncertainties, which are simply RU = $\frac{\text{U observed}}{\text{U maximum}}$. The formula for RU is also presented in Appendix A. In the present study one set of items, the mathematical item type Quantitative Comparisons, was basically a four-choice model. The other item types were five-choice. The use of RU, therefore, was required.

For any set of response categories, Shannon's function is at a maximum for a rectangular distribution, so that a high RU value tests the extent to which there are equal proportions of responses to the various categories.

Both indices are necessary and useful in the evaluation of items. To underscore these points, three hypothetical items with extreme characteristics are presented in Figures 1-3. These items illustrate the duality of the requirement of suitable values for both indices and at the same time show how r_{pw} is in theory independent of the biserial r. In each case it is assumed that the item has a basic biserial correlation under dichotomous scoring of .50, and an item difficulty of .50. That is, 50% of the group succeeds, and their mean standard score (z-score) is approximately +.40.

Figure 1 presents an item for which RU = 1.0, so that the potential information is at a maximum. This is indicated by the rectangular distribution of percentages across the wrong answers (lower line). The polyserial, however, is 0.00, indicated by the absence of slope for the upper line (mean z-scores) in the region of the wrong answers.

Figure 2 presents an optimal item, one for which RU = 1.0, r_{pw} = 1.0. In this optimal item, two distracters have positive mean z-scores. In fact, the best distracter, if scored as the key, would have a biserial of about .57, and a difficulty of .10. The worst distracter captures virtually all of the low-scoring persons, and has a biserial approaching -1.0 if scored alone. Clearly, this is no ordinary item.

Figure 3 presents an item with a high polyserial but a low value of RU. The five non-right responses (four Wrong answers and an Omit response) account for 40%, 2.5%, 2.5%, 2.5%, and 2.5% of the total group, respectively. While the polyserial is high, the potential usefulness of the item is diminished by the failure to provide more differentiations, somewhat as an extremely easy or difficult item is of limited usefulness in conventional scoring. Figure 3 underscores the need for appropriate values on both indices of item potential.

Figures 4 and 5 describe actual SAT items. Figure 4 is a Mathematical Reasoning item with a difficulty of .36 and a biserial of .35. It has a fairly high RU (.91) and one of the higher values of r_{pw} (.44) among the 300-odd items considered in the study. Figure 5 is a verbal antonym item with a difficulty of .55 and a biserial of .57. It has a very high RU (.99) but a relatively low r_{pw} (.18) indicates only a limited measurement potential for weighting.

Studies of SAT item types

The present study could make only a limited assessment of the utility of the indices. No weighting, for example, was possible. It was decided to survey the frequency distributions for both r_{pw} and RU in four major item types used in the College Board Scholastic Aptitude Test. In all, 302 items were studied, derived from 4 recent forms of the SAT: 51 reading comprehension (RC), 100 antonyms (VA), 91 mathematical reasoning (MR), and 60 mathematical quantitative comparisons items (MQ). With the exception of MQ items, which are four-choice, all items offered five choices. MQ items ask for a judgment as to the relative magnitude of two quantities, one in column A and one in column B. For example:

A	B
x^3	x^2

The four answers to any MQ item are A (A larger than B), B (B larger than A), C (A = B) and D (not enough information to decide).

The distributions of values of r_{pw} for the four item types are presented in Table 1. The means range from .19 for VA to .23 for MR. The two mathematical item types tend to have higher values than the two verbal types. This pattern is also reflected in the biserial correlations for the items.

In general, the results reflect only a modest potential for weighting on the part of the average SAT item. The levels of polyserial correlation are typically lower than the level of biserial correlation. There are individual items with better than average promise, however. If a test was composed entirely of items with a polyserial greater than .25, it might be

a significantly better candidate for differential weighting than the average SAT test. In particular, the replication of studies such as Hendrickson's (1970) study of the SAT, but with the introduction of items of higher average quality, could shed light on the extent to which her findings were limited by the average quality of the items.

The distributions of values of RU are presented in Table 2. These distributions are all negatively skewed, indicating that a fairly high proportion of the possible information, in Shannon's sense, was attained. The scale properties of Shannon's measure, or of Pike and Flaugher's variation, may not be well adapted for test construction, and a transformation of some type may be necessary if the range of greatest density is in fact of critical importance to test developers.

The mathematical quantitative comparisons items (MQ) show markedly lower average values of RU. This would indicate that the three wrong answers in this item type are less likely to be equally attractive than in a conventional item. This result seems plausible, given the nature of the item decision space, in which the responses are logically related, so that A is the opposite of B and vice versa. The implication of the lower average RU value is that this item type would possibly be less useful for weighting than the others. Or, put more correctly, that a larger pool of MQ items might need to be prepared in order to achieve an average value of RU equal to that of the other item types.

Tables 3 and 4 present the distributions of item difficulties and item biserials for the four item types. They indicate a basic similarity in the

average values of the items, with the two mathematical items being both easier and more highly correlated, on the average.

The values of r_{pw} and RU are in theory independent of item difficulty and item biserial correlation. However, such theoretical characteristics are often better confirmed than assumed. For example, in theory biserial r is independent of item difficulty, while in practice there is a marked relationship. The effect of the greater amount of more random guessing behavior on difficult items depresses the biserial correlation. Table 5 presents the intercorrelations of the variables, both within item type and across all items. There are modest associations between the proposed indices and the traditional ones in some cases. Thus, r_{pw} correlates -.27 with r_{bis} for MQ items, and RU correlates -.28 with p^+ for RC items. In general, however, the new measures are sufficiently independent from the old to warrant a consideration of their unique contribution.

There is quite a lot of variation from item type to item type. These samples are quite small, however, The most stable relationship is the correlation between the two "traditional" measures: .48 across all items. As remarked earlier, the biserial correlation tends to be lower for difficult items. The negative correlation for RU and p^+ indicates that the higher average p^+ for the MQ items may partially explain their lower average RU values.

Discussion

The basic premise of the study seems substantiated: items do exhibit varying levels of the two indices and in ways that seem consistent with the assumption that higher values of the indices will contribute to weighting. That the internal structure of items is reflected in the varying values of the indices is intuitively supported by the graphs in Figures 1-5, which show the variation in potential for a variety of hypothetical and empirical items.

The full interpretation of the new statistics is not possible in the absence of more experience with them. The relation of item properties to score properties is fairly complex for dichotomously-scored items, and may be anticipated to be even more complex for differentially weighted material. The degree to which tests of different average level of polyserial show different score properties when weighted is at present largely conjectural. Theoretical expication of the relationships is possible, particularly if the kinds of practical measures suggested here can be linked to the more powerful latent trait models such as those of Thissen (1976). But the theoretical work is likely to be difficult and to demand a large body of assumption not often well met in practice.

Alternate approaches would appear to lie in the direction of Swineford's (1959) study of the relationship of item difficulty and discrimination indices to test properties. In her study, a large number of test forms of varying average item characteristics were studied, and a set of linear relationships established between such item data and the resultant score

properties. Given the apparently more limited ability of test developers to achieve high levels of polyserial correlation, a more fruitful approach may be in a Monte Carlo inquiry relying upon computer simulation. In either case, whether "real world" tests are possible or computer analogs are generated, the resultant equations of relationship may yield tools for test developers at a level of practicality not achievable through theory.

Although complete interpretation is not possible, it seems appropriate to conclude that the average promise for weighting of the typical current SAT item is not great. Individual items may have better characteristics, but the general level is not such that the additional work of differential scoring would be worthwhile. The problem seems to lie more in the failure to achieve ordered means among the wrong answers, than in a failure to have a spread of responses across several distracters. That is, the polyserials seem low but the RU information functions do not seem low.

Attention should be called to the lower average RU values for the QC item type. This item type would seem to offer many fewer successful items for differential weighting, a finding which seems consistent with the internal logical structural relationships. Findings of relationships between the solution-process characteristics of different item types and their potential for weighting could significantly advance work in this area by increasing the control which test developers have.

The study was frankly exploratory and not designed for rigorous hypothesis testing. Within these limits, the conclusions would appear to be:

1) the proposed indices of item correlation (wrong-answer polyserial) and distribution (Relative Uncertainty) show an ability to reflect desirable characteristics for weighting

2) The average potential of four SAT item types for weighting is not great

3) this limited potential for weighting derives more from a failure to generate ordered means than a sufficiently broad distribution

4) The RU values of the QC item type are lower than those of the other item types

Proposed additional work would cover such activities as a) evaluating the new indices in comparison with other possible indices such as the point-multiserial, b) developing theoretical lin s to models such as Thissen's or other latent-trait, item-characteristic curve models, and c) exploring broad-gauge empirical inquiries along the line of Swineford's, using either real-world data or computer simulation. In view of the long history of fruitless attempts to demonstrate useful outcomes for differential weighting, it is perhaps presumptuous to urge continued attention to the area. But the topic of item potential for weighting seems largely unexplored, and attention to this line of approach seems indicated.

References

Aiken, L.R. Effect on test score variance of differential weighting of item responses. Psychological Reports, 1967, 21, 585-590.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. Part 5 (Chs. 17-20, pp. 395-479) in F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Cleary, T.A. An individual difference model for multiple regression. Psychometrika, 1966, 215-224.

Das Gupta, S. Point Biserial Correlation Coefficient and its Generalization. Psychometrika, 1960, 393-408.

Davis, F.B. Estimation and use of scoring weights for each choice in multiple choice test items. Educational and Psychological Measurement, 1959, 19, 291-298.

Davis, F.B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

Fruchter, B. Differences in factor content of right and wrong scores. Psychometrika, 1953, 257-265.

Gronlund, N.E. Measurement and Evaluation in Teaching (3rd. Ed.) New York: MacMillan, 1976.

Guilford, J.P. Psychometric methods, New York: McGraw-Hill, 1954.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Guttman, L. An outline of the statistical theory of prediction. In Paul Horst, The prediction of personal adjustment. New York: Social Science Research Council, 1941.

Hambleton, R., and Cook, L. Latent trait models and their use in analyzing educational test data. Journal of Educational Measurement, 1977, 75-96.

Hendrickson, Gerry F. The effect of differential option weighting on multiple-choice objective tests. Journal of Educational Measurement, 1971, 291-296.

Henrysson, S. Gathering, Analyzing and Using Data on Test Items. in Thorndike, R.L. (ed.) Educational Measurement, Washington, D. C., American Council on Education, 1970.

Jacobs, P.I., & Vandeventer, M. Information in wrong responses. Research Bulletin, 68-25, Educational Testing Service, June 1968.

Jaspen, W. Serial correlation. Psychometrika, 1946, 23-30.

Kelley, T.L. Statistical method. New York: Macmillan, 1923.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 1954, 459-472.

Odell, C.W. Further data concerning the effect of weighting exercises in new-type examinations. Journal of Educational Psychology, 1931, 700-704.

Pike, L., & Flaugher, R. Assessing the Meaningfulness of Group Responses to Multiple Choice Test items. APA Proceedings, 1970, 101-102.

Sabers, D.L., & White, G.W. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. Journal of Educational Measurement, 1969, 93-96.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, No. 17.

Serlin, R.C., & Kaiser, H.S. A method for increasing the reliability of a short multiple-choice test. Educational and Psychological Measurement, 1978, 337-340.

Shannon, C.L. A mathematical theory of communication. Bell System Technical Journal, 1948, 379-423; 623-656.

Stalnaker, J.M. Weighting questions in the essay-type examination. Journal of Educational Psychology, 1938, 481-490.

Stanley, J., & Wang, M.D. Differential weighting. New York: College Entrance Examination Board, 1968.

Swineford, F. Some relations between test scores and item statistics. Journal of Educational Psychology, 1959, 26-36.

Thissen, D.M. Information in Wrong Responses to the Raven Progressive Matrices. Journal of Educational Measurement. 1976, 201-214.

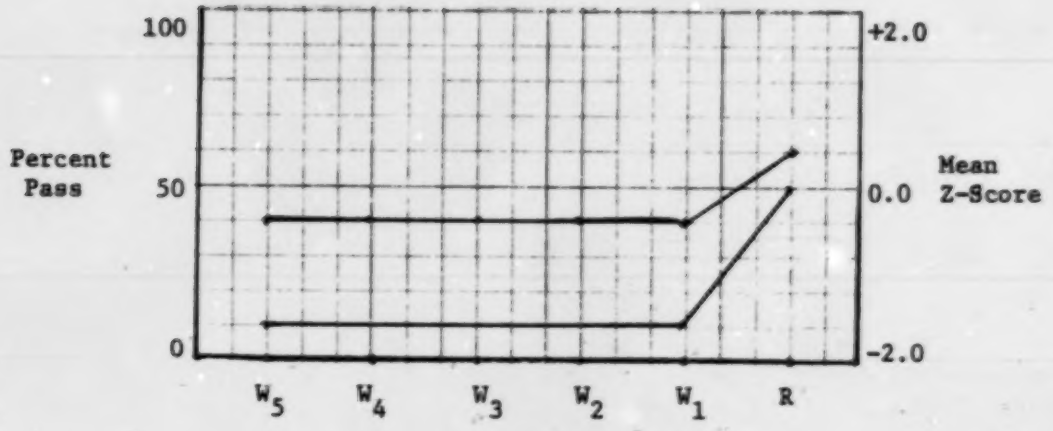
Thurstone, L.L. A scoring method for mental tests. Psychological Bulletin, 1919, 235-240.

Wang, M.D., & Stanley, J. Differential weighting: a review of methods and empirical studies. Review of Educational Research, 1970, 663-705.

West, P.V. The significance of weighted scores. Journal of Educational Psychology, 1924, 302-308.

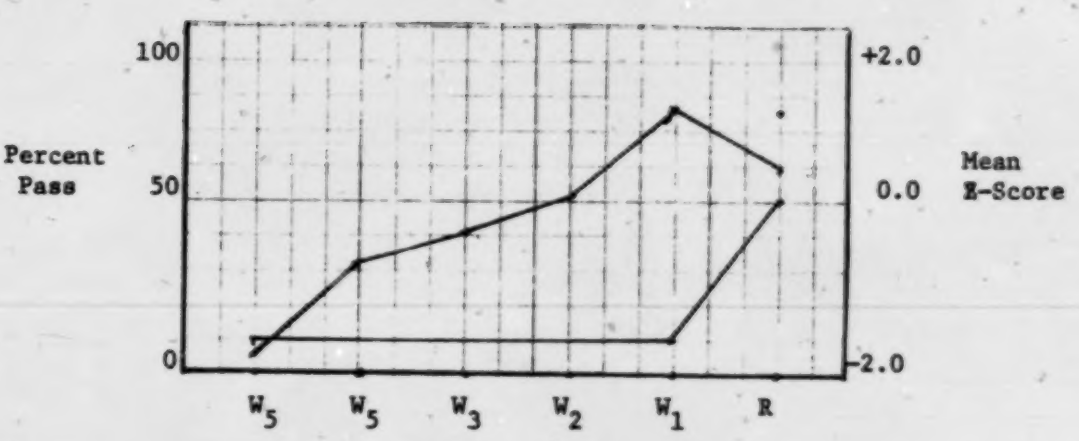
Wherry, R.J., & Taylor, E.K. The relation of multiserial eta to other measures of correlation. Psychometrika, 1946, 155-162.

Wilks, S.S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 23-40.



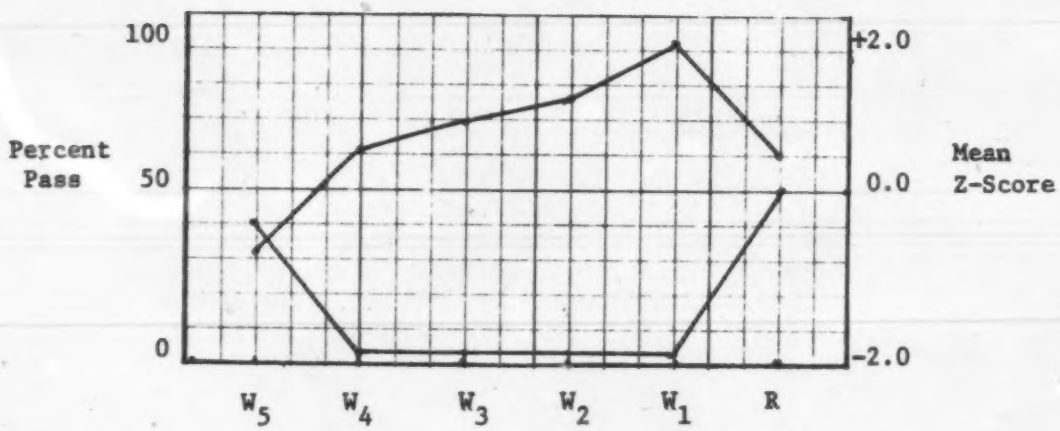
Percent pass (lower line) and mean z-score (upper line) for Wrong Responses (W) and Right Answer (R) item characteristics r_{pw} = 0.00, RU = 1.0, p^+ = .50, r_{bis} = .50

FIG. 1



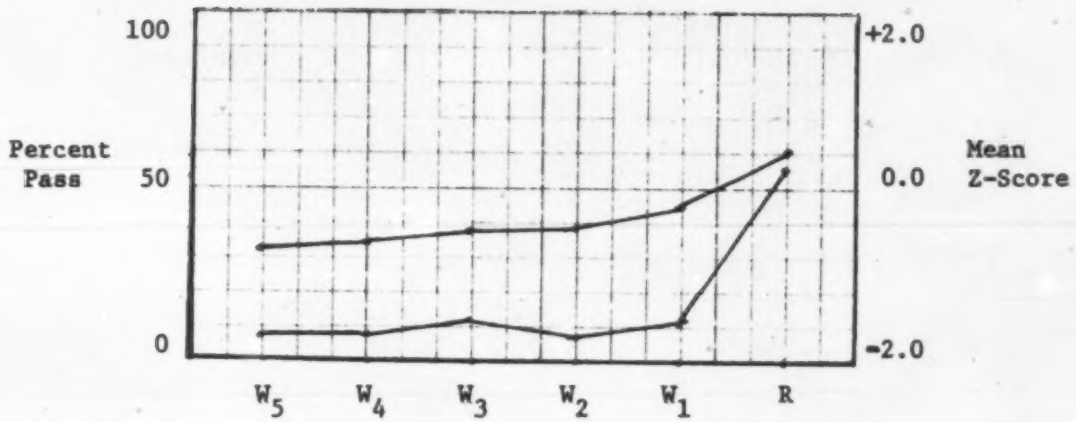
Percent pass (lower line) and mean z-score (upper line) for Wrong Responses (W) and Right Answer (R) item characteristics r_{pw} = 1.0, RU = 1.0, p^+ = .50, r_{bis} = .50

FIG. 2



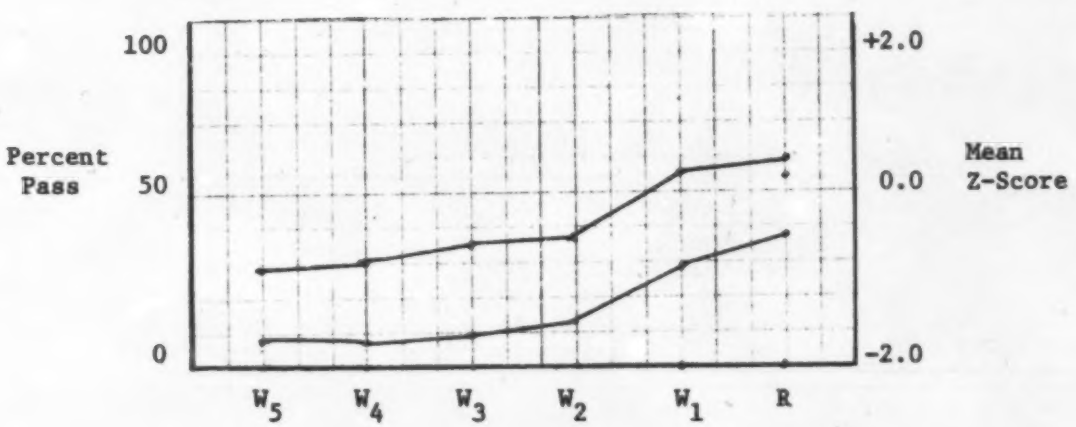
Percent pass (lower line) and mean z-score (upper line) for Wrong Responses (W) and Right Answer (R) item characteristics r_{pw} = 0.00, RU = .48, p^{+} = .50, r_{bis} = .50

FIG. 3



Percent pass (lower line) and mean z-score (upper line) for Wrong Responses (W) and Right Answer (R) item characteristics r_{pw} = .18, RU = .99, p^+ = .55, r_{bis} = .57

FIG. 4



Percent pass (lower line) and mean z-score (upper line) for Wrong Responses (W) and Right Answer (R) item characteristics r_{pw} = .44, RU = .91, p^+ = .36, r_{bis} = .35

FIG. 5

TABLE 1

Distributions of r_{pw} for Four Item Types

r_{pw}	RC	VA	MR	MQ
56-60				1
51-55		1		
46-50		1	1	2
41-45	1	1	6	2
36-40	1	2	3	1
31-35	1	7	5	3
26-30	9	6	10	8
21-25	4	8	22	13
16-20	21	39	26	11
11-15	8	25	14	13
06-10	5	8	4	5
01-05	1	2		1
N	51	100	91	60
Mean	.19	.19	.23	.22
σ	.07	.09	.09	.11

TABLE 2

Distributions of RU for Four Item Types

RU	RC	VA	MR	MQ
96-1.00	8	22	19	9
91- 95	10	42	32	10
86- 90	8	19	16	11
81- 85	7	7	5	5
76- 80	8	5	8	4
71- 75	5	2	7	7
66- 70	1			2
61- 65	1	1	2	5
56- 60	1			2
51- 55	1	1		1
46- 50				
41- 45			1	1
36- 40	1		1	
31- 35				1
26- 30		1		2
N	51	100	91	60
Mean	.84	.90	.88	.79
σ	.12	.09	.11	.15

TABLE 3

Distributions of Item Difficulty (p) for Four Item Types

p		RC	VA	MR	MQ
91-1.00		1	3	2	3
81- 90		5	17	8	9
71- 80		7	24	22	17
61- 70		10	7	19	11
57- 60		11	7	16	8
41- 50		6	2	13	5
31- 40		6	6	9	
21- 30		4	14	2	4
11- 20		1	17		3
1- 10			3		
	N	51	100	91	60
	Mean	.57	.53	.62	.64
	σ	.19	.28	.16	.20

TABLE 4

Distributions of Biserial Correlations (r_{bis}) for Four Item Types

r_{bis}		RC	VA	MR	MQ
71- 80			2	8	3
61- 70		4	18	35	16
51- 60		13	25	29	20
41- 50		24	29	13	15
31- 40		8	18	5	3
21- 30		1	6	1	1
11- 20		1	2		1
	N	51	100	91	60
	Mean	.47	.48	.59	.55
	σ	.09	.13	.10	.13

TABLE 5

Intercorrelations of r_{pw}, RU, p^+ and r_{bis} for the four item types and for total items

	RU	p^+	r_{bis}
r_{pw}	.03	.04	-.17
RU		-.28	.38
p^+			.37

RC items

N = 51

	RU	p^+	r_{bis}
r_{pw}	-.37	.26	-.06
RU		-.16	.01
p^+			.68

VA items

N = 100

	RU	p^+	r_{bis}
r_{pw}	-.15	.15	-.04
RU		-.22	-.15
p^+			.16

MR items

N - 91

	RU	p^+	r_{bis}
r_{pw}	-.14	.22	-.27
RU		.04	.05
p^+			.32

MQ items

N = 60

	RU	p^+	r_{bis}
r_{pw}	-.18	.22	-.04
RU		-.19	-.02
p^+			.48

Total items

N = 302

APPENDIX A

Jaspen's formula for serial correlation may be stated as follows:

$$r_p = \frac{\sum_i^k \bar{z}_i \, (y_i - y_{i-1})}{\sigma_y \sum_i^k \frac{(y_i - y_{i-1})^2}{p_i}}$$

where $\bar{z}_i$ = mean z-score in the i-th interval of the subdivided unit normal curve.

y_i = ordinate under the normal curve at lower (left-hand) end of the i-th interval.

y_{i-1} = ordinate under the normal curve at upper (right-hand) end of i-th interval.

p_i = proportion of area under unit normal curve in i-th segment.

σ_y = standard deviation of the fully-distributed variable.

k = the number of categories.

Pike and Flaugher's index is:

$$RU = \frac{\sum_i^k - p_i \log_2 p_i}{\log_2 k}$$

where p_i = proportion of the distribution in response category i.

k = the number of categories.